# Statistical methods for the assessment of clusters discovered in bio-molecular data.

Giorgio Valentini

DSI – Dipartimento di Scienze dell'Informazione,
Università degli Studi di Milano
e-mail: valentini@dsi.unimi.it

## Abstract

The assessment of the reliability of clusters discovered in bio-molecular data is a central issue in several bioinformatics problems, ranging from the definition of new taxonomies of malignancies based on bio-molecular data, to the validation of clusters of co-regulated or co-expressed genes, or the discovery of functional relationships from protein-protein interaction data.
Recently, several methods based on the concept of stability have been proposed to estimate the reliability and the "optimal" number of clusters. In this conceptual framework multiple clusterings are obtained by introducing perturbations into the original data, and a clustering is considered reliable if it is approximately maintained across multiple perturbations. Different procedures have been introduced to randomly perturb the data, ranging from bootstrapping techniques, to noise injection into the data or random projections into lower dimensional subspaces.
Usually, stability-based methods provide only a score or a measure of the reliability of the discovered clusters, without any assessment of the statistical significance of the clustering solutions; moreover they are not able to directly detect multiple structures (e.g. hierarchical structures) simultaneously present in the data. Recently we proposed a chi squared-based statistical test and a distribution-free test based on the classical Bernstein inequality, showing that stability-based methods can be successfully applied to the assessment of the reliability of clusterings, as well as to discover multiple structures underlying complex bio-molecular data.

## Introduction

Several tasks related to the analysis of bio-molecular data require the development and application of unsupervised clustering techniques [1,2]. Unfortunately, clustering algorithms may find structure in the data, even when no structure is present instead. Moreover, even if we choose an appropriate clustering algorithm for the given data, we need to assess the reliability of the discovered clusters, and to solve the model order selection problem, that is the proper selection of the "natural" number of clusters underlying the data [3,4]. From a machine learning standpoint, this is an intrinsically "ill-posed" problem, since in unsupervised learning we lack an external objective criterion. From a biological standpoint, in many cases we have no sufficient biological knowledge to "a priori" evaluate both the number of clusters (e.g. the number of biologically distinct tumor classes), as well as the validity of the discovered clusters (e.g. the reliability of new discovered tumor classes) [5].
Recently, several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters [3,6]: multiple clusterings are obtained by introducing perturbations into the original data, and a clustering is considered reliable if it is approximately maintained across multiple perturbations. Different procedures may be applied

to randomly perturb the data, ranging from bootstrapping techniques, to noise injection into the data or random projections into lower dimensional subspaces [7,8]. Stability indices, obtained from the distribution of the similarities computed between multiple perturbed clusterings, are usually applied to quantitatively estimate the reliability of the clusterings, and statistical tests have been proposed to assess the significance of clustering solutions [3,6,7,8,9]. In this contribution we briefly review the characteristics of stability based methods and two statitistical tests recently proposed to assess the significance of clustering solutions and to discover multiple structures underlying high-dimensional bio-molecular data.

**Stability based methods**

In this conceptual framework multiple clusterings are obtained by introducing perturbations into the original data, and a clustering is considered reliable if it is 'stable', that is if it is approximately maintained across multiple perturbations. The main logical steps of a high-level algorithm for a stability-based evaluation of clustering solutions may be summarized as follows:

1. Perform multiple random perturbations of the data.
2. Supply the perturbed data to a given clustering algorithm.
3. Apply a given clustering similarity measure to multiple pairs of k-clusterings obtained according to steps (1) and (2).
4. Use appropriate stability indices, based on the similarity measures applied at step (3), to score the reliability of a given k-clustering.
5. Repeat steps (1) to (4) for different numbers k of clusters and elect the most stable clustering(s) as the most reliable.

Random perturbations (step 1) may be realized through random noise injection into the data [7] or by subsampling techniques, or random projections into lower dimensional subspaces [3,9]. The similarity between clusterings (step 3) can be estimated by means of classical measures (e.g. the Jaccard index) [10]. The stability indices for model order selection proposed in the literature (step 4) can be schematically divided into ones that use statistics of the similarity measures [7,8,9] and others that exploit their overall distribution [3]. The last step is usually approached by choosing the best scored clustering (according to the chosen stability index), but a major problem is represented by the estimate of the statistical significance of the discovered solutions.

**Statistical tests to assess the significance of clusters discovered in bio-molecular data**

In [3] a $\chi^2$-based test has been proposed to assess the significance of clustering solutions and to discover multiple structures in bio-molecular data. Here for structure we mean a *k-clustering*, that is a clustering composed by *k* clusters. By this method we can test the following hypotheses:

*Null hypothesis (Ho): all the the considered set of k-clustering are equally reliable. Alternative hypothesis (Ha): the considered set of k-clustering are not equally reliable.*

An *iterative procedure* that exploits the ordering of the stability indices to detect the significant number(s) of clusterings is proposed:

1. Consider the ordered vector (from the most to the least reliable) $\xi = (\xi_{p(1)}, \xi_{p(2)}, ..., \xi_{p(H)})$ of the stability indices

2. Repeat the $\chi^2$-based test until no significant difference is detected or the only remaining clustering is *p(1)* (the top-ranked one). At each iteration, if a significant difference is detected, remove the bottom-ranked clustering from $\xi$.

*Output*: set of the remaining (top sorted) *k*-clusterings that correspond to the set of the estimate reliable number of clusters (at $\alpha$ significance level). For more details, please see [3].

This statistical test presents two main drawbacks: a) a priori assumptions about the distribution of the similarity values needed to estimate the reliability of the obtained clusterings b) test results depend on the choice of user-defined parameters.

To overcome these problems, an alternative approach based on the classical *Bernstein inequality* [11] have been recently proposed [12]. Indeed by this method no assumptions about the distribution of the similarity values are made, and no requirements of any user-defined additional parameters are needed, thus assuring a reliable application to a large range of bioinformatics problems.

**An experimental comparison of the $\chi^2$ and *Bernstein*-based test for the discovery of multiple structures in gene expression data**

To show the effectiveness of the proposed methods we apply them to the analysis of the *Leukemia* DNA microarray data set [13]. This gene expression data set is composed by a group of 25 acute myeloid leukemia (AML) samples and another group of 47 acute lymphoblastic leukemia (ALL) samples, that can be subdivided into 38 B-Cell and 9 T-Cell subgroups, resulting in a two-level hierarchical structure.

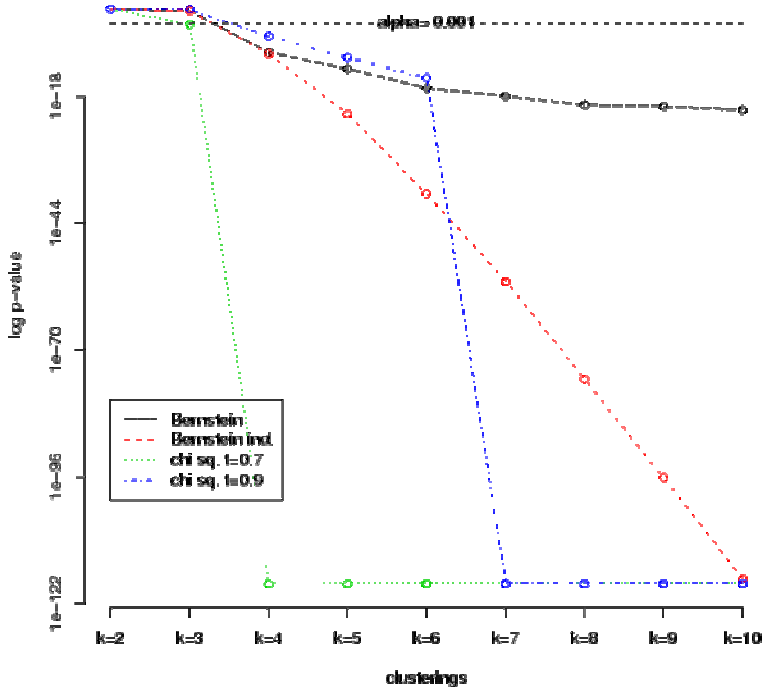


Figure 1: K-means clustering: log *p-values* computed for $\chi^2$-based and *Bernstein*-based statistical tests. Ordinate: log *p-value*; abscissa: number of clusters sorted according the computed stability indices.

Results are summarized in Figure 1: a straight horizontal dashed line represents a significance level $\alpha = 0.001$: k-clusterings above the dashed line are significant, that is their reliability significantly differ from the *k*-clusterings below the dashed horizontal line. Note that both tests consider 2 and 3 clusterings significantly more reliable than the others, according to the biological characteristics of the data; *Berstein ind*. assumes independence between the random variables that represent the stability indices. Nevertheless, the *Bernstein* test, due to its more general assumptions is less selective (in the sense that it may consider reliable a larger number of *k*-clusterings) than the $\chi^2$-based test that make assumptions about the distribution of the similarity values. This is confirmed by the fact that *Bernstein* p-values decrease more slowly with respect to the $\chi^2$ test, thus resulting in a better sensitivity to multiple structures present in the data. The main drawback of this behaviour is the larger probability of false positives.

## Conclusions

We briefly presented stability based methods to estimate the reliability of clusterings discovered in bio-molecular data. These methods, if jointly used with statistical tests specifically designed to discover multiple structures (i.e. *k*-clusterings), can be successfully applied to assess the statistical significance of clusterings in complex bio-molecular data.

## References


[1] Gasch P, Eisen M: Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology* 2002, 3(11).

[2] Kaplan N, Friedlich M, Fromer M, Linial M: A functional hierarchical organization of the protein sequence space. *BMC Bioinformatics* 2004, 5

[3] Bertoni A, Valentini G: Model order selection for bio-molecular data clustering. *BMC Bioinformatics* 2007, 8(Suppl.3).

[4] Garge N, Page G, Sprague A, Gorman B, Allison D: Reproducible Clusters from Microarray Research: Whither? *BMC Bioinformatics* 2005, 6(Suppl2).

[5] Alizadeh A, Ross D, Perou C, van de Rijn M: Towards a novel classification of human malignancies based on gene expression. *J. Pathol.* 2001, 195:41-52.

[6] Lange T, Roth V, Braun M, Buhmann J: Stability-based Validation of Clustering Solutions. *Neural Computation* 2004, 16:1299-1323.

[7] McShane L, Radmacher D, Freidlin B, Yu R, Li M, Simon R: Method for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* 2002, 18(11):1462-1469.

[8] Bertoni A, Valentini G: Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses. *Artificial Intelligence in Medicine* 2006, 37(2):85-109

[9] Smolkin M, Gosh D: Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics* 2003, 36(4).

[10] Jain, A.K. and Dubes, R.C. *Algorithms for Clustering Data*, (1988) , NJ Prentice Hall.

[11] Hoeffding W: Probability inequalities for sums of independent random variables. *J. Amer. Statist. Assoc.* 1963, 58:13-30.

[12] A.Bertoni, G.Valentini, A statistical test based on the Bernstein inequality to discover multi-level structures in bio-molecular data BITS 2007, *Bioinformatics Italian Society Meeting*, Napoli, Italy, 2007.

[13] Golub T, et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 1999, 286:531-537.